

Detection of Spatiotemporally Coherent Rainfall Anomalies Using Markov Random Fields

Adway Mitra, Ashwin K. Seshadri

March 28, 2017

Abstract

Precipitation is a large-scale, spatio-temporally heterogeneous phenomenon, with frequent anomalies exhibiting unusually high or low values. We use Markov Random Fields (MRFs) to detect spatio-temporally coherent anomalies in gridded annual rainfall data across India from 1901-2005. MRFs are undirected graphical models where each node is associated with a {location, year} pair, with edges connecting nodes representing adjacent locations or years. Some nodes represent observations of precipitation, while the rest represent unobserved (*latent*) states that can take one of three values: high/low/normal. The MRF represents a probability distribution over the variables, using *node potential* and *edge potential* functions defined on nodes and edges of the graph. Optimal values of latent state variables are estimated by maximizing the posterior probability of the observations, using Gibbs sampling. Edge potentials enforce spatial and temporal coherence, and node potentials influence threshold for anomalies by affecting the prior probabilities of the states. The model can be tuned to recover anomalies detected by threshold-based methods. The competing influences of spatial and temporal coherence can be adjusted through edge potentials.

We study spatio-temporal properties of rainfall anomalies discovered by this method, using suitable measures. We identify nonstationarities in occurrence of positive and negative anomalies between the first and second halves of the 20th century. We find that between these periods, there has been decrease in rainfall during June-September (JJAS) and an increase during other months. These effects are highlighted prominently in the statistics of anomalies. Properties of anomalies learnt from this approach could present tests of regional-scale rainfall simulations by climate models and statistical simulators.

1 Introduction

In many parts of the world, such as India, rainfall plays an important role in the economy and the well-being of millions of people. Consequently, anomalies of rainfall can have very significant effects. It is known that low annual rainfall

has an adverse effect on India’s GDP [2]. In climate science, “anomaly” of a climatic variable (like precipitation at a particular location) is defined quantitatively, as the magnitude of deviation from its climatological value, averaged over many years. But in this work, we will use the term “anomaly” qualitatively, indicating deviation not only from its climatological value, but also with respect to its spatial/temporal neighbors. Anomalies can occur at different spatial and temporal scales, and their occurrence is heterogeneous (the statistics are location-dependent) and anisotropic (not uniform in all directions). Generally, both positive and negative anomalies are present within the same year. The more consequential anomalies are the ones with significant spatiotemporal extent, and therefore it is important to identify them. Another factor is that with climate change, the frequency of rainfall extremes may increase, along with changes in the spatial pattern of rainfall[5].

To understand past and future changes, scientists rely on global climate models (GCMs) which simulate global climatic variables including rainfall. Algorithms are necessary for analyzing large-scale simulations as well as observational data procured from sensors, and such analyses should include detecting and summarizing statistics of rainfall anomalies ([8, 6, 7]). Such analysis cannot be done manually because of the large and growing volume of data and simulation results, raising the need for automated procedures.

Automating anomaly detection is challenging, because anomalies are inherently subjective, depending on definition and detection threshold [9]. Spatiotemporal anomaly detection is considered an important research area in Data Science[15]. The simplest approach to anomaly detection is based on a predefined threshold, relative to statistics of the corresponding variable. With rainfall, one might consider the time-series of annual mean rainfall at each grid location, estimate its mean and variance, and identify years departing significantly from the mean. However, accounting for effects of spatiotemporal neighbours is important for detection [11], and the aforementioned location-wise threshold-based approach cannot do this.

Furthermore the anomaly detection problem is broader, especially when the anomaly is conceived as a conceptual or abstract quantity represented by a state variable that cannot be directly observed or measured and must be inferred indirectly. Here we consider anomalies in rainfall as a *latent variable*, as often done in statistical modelling [13] including spatiotemporal modelling [14]. Such latent (i.e. unobserved) states are best estimated through probabilistic methods[13, 12]. We associate a latent state variable with each spatiotemporal location, i.e. each combination of grid-point and year. These variables are modelled to be spatiotemporally coherent through parameters of the MRF.

In this work we estimate these latent variables as the maximum posterior (MAP) solution of a Markov Random Field (MRF). MRFs are undirected random graphical models satisfying Markov properties, and are generally used to model joint distributions of several variables[16]. Given a likelihood model of the data conditional on the states of this graph, the posterior density and correspondingly the MAP solution of this graph can be estimated.

The nodes in our MRF correspond to unique combinations of {spatial,temporal}

locations. Each node corresponds to either observed volume of rainfall at that combination of location (grid-point) and time (year), or the corresponding latent variable describing possible occurrence of an anomaly state. Each latent state node has three values: 1 (positive anomaly), 2 (negative anomaly) or 3 (normal). Two latent state nodes have an edge between them, i.e. are neighbours in the MRF, if they are adjacent in either space or time. Each observation node is connected only to the latent state variable node for that spatiotemporal location.

MRFs are defined using "potential functions" for nodes and edges of the graph, which encode interactions between neighbouring variables. In our application, these functions influence the spatial and temporal coherence of the state variables.

The local Markov properties inherent to MRFs imply that, for any node, its state is conditionally independent of all other nodes except its neighbours. This property of localness or memorylessness with respect to conditional distributions is quite familiar in many contexts; but with anomaly detection of spatiotemporal fields with large-scale character it has an additional effect. Specifically, where large-scale teleconnections induce correlated anomalies across distance, the Markov property ensures that these are counted as different anomalies.

While there are other approaches to anomaly detection [9] including in case of spatiotemporal anomalies [15], here we use MRFs for studying coherent rainfall anomalies. MRFs themselves have been used in similar applications involving geospatial fields, including rainfall. Fu et al. (2012) [4] have used MRFs to detect coherent droughts of the last century, and find that their procedure can identify well-known droughts around the world. Theirs appears to be the first formulation of the rainfall anomaly detection problem in terms of MRFs. That work used integer programming to identify the MAP configuration of the latent states [4]. However, integer programming is very slow, increasing exponentially in the size of the problem, thereby necessitating probabilistic inference techniques [13, 12].

The present paper is partly motivated by the aforementioned work [4]. We focus on grid-level annual rainfall over India, but our method is general enough to work on rainfall data at any spatial and temporal resolution.

Like [4] we use Markov Random Fields, but an important difference is that both positive and negative anomalies are considered, so that the latent variable in each node is in one of three states (positive, negative, normal). In addition the relation between anomalies at small scales (grid-wise) and large scale (all-India spatial mean) is explicitly modelled.

Furthermore, in contrast to [4] we use Gibbs sampling to infer the latent variables. Gibbs sampling works by creating a Markov chain whose stationary distribution is the distribution we seek, and then carrying out a random walk on this Markov chain ([17, 18]). Gibbs sampling has been used previously in estimating MRFs (e.g. Rue[10]), and here we illustrate its usefulness in estimating latent states corresponding to large and heterogeneous geospatial fields such as rainfall. Compared to integer programming, the Gibbs sampling algorithm is computationally inexpensive.

Based on the inferred latent states we identify spatiotemporally coherent anomalies, and quantify their properties. Effects of enforcing spatial and temporal coherence on the resulting anomalies are examined, and sensitivity to parameters is studied. We compare the spatial extents of positive and negative anomalies. There is an inherent trade-off between spatial and temporal extents of anomalies in any procedure, originating in the values of parameters enforcing spatial and temporal coherence. Furthermore, even for any fixed set of parameters, there is variability in the spatial and temporal sizes of the detected anomalies. Both of these effects are examined.

Finally we consider temporal nonstationarity in the frequency and extents of anomalies, by summarizing differences between anomalies detected for 1901-1950 and 1951-2000. In the second half of the century there has been a shift in rainfall from the core monsoon season (June-September) to other months, and significant increases are observed in the pre-monsoon (April-May) and post-monsoon (October-November). This shift is present in the month-wise distribution of accumulated rainfall, but the effects are more prominent when anomalies are considered. Furthermore, the analysis of anomalies reveals additional aspects of this change.

2 Methodology

2.1 Definitions and Notation

We consider S locations and T years, and spatiotemporal observations Y_{st} of a geophysical variable such as annual-mean rainfall. Then s indexes location and t indexes time. We consider each location s to be in one of three possible *states* in any year t - high (1), low (2) or normal (3). This follows the conventional classification of rainfall-years as positive anomaly (excess rainfall), negative anomaly (deficient rainfall), or normal. The state is represented by a latent discrete variable Z_{st} taking one of 3 values. A goal of anomaly detection is to estimate these latent variables, from which anomalies can be identified [9].

2.2 Location-wise Analysis

A naive solution is to treat the time-series at each location individually. For each time-series we compute mean μ_s and standard deviation σ_s . We then set $Z_{st} = 1$ (high) for those years where $Y_{st} \geq \mu_s + \sigma_s$, $Z_{st} = 2$ (low) for those years where $Y_{st} \leq \mu_s - \sigma_s$, and $Z_{st} = 3$ (normal) for all other years. We call this method Location-Wise Analysis (LWA), since it treats each location independently without considering the state of its neighbours. Corresponding assignments to the latent variables by this method are denoted as $Z0$. The limitation of this approach is of course its neglect of spatial coherence in the latent variable. For example an individual location may be in a certain mode, while all its neighbours are in a different mode in the same year. Isolated anomalies need not be spurious, but spatially or temporally extended anomalies

are more consequential.

Detecting extended anomalies requires a different lens, one inducing spatial or temporal coherence in the Z_{st} -variables. We seek to discover spatial and temporal clusters within which Z -values are the same. An alternate approach might be to undertake location wise analysis, after having smoothed data onto a coarser grid. This enlarges the scales of interest, but involves loss of information. It also does not permit anomalies at multiple scales, or naturally accommodate spatial heterogeneity or anisotropy in anomalies.

2.3 Modeling by Markov Random Fields

To address this shortcoming, we take the approach that assigns probabilities to different configurations of latent Z -variables, with higher weights to configurations where Z -assignments are spatially or temporally coherent. This is achieved by modelling the latent variable as an MRF, along the lines of the drought discovery technique in [4].

In addition to grid-wise latent states, these can also be defined for the all-India mean, relative to its corresponding distribution across years. The Indian Meteorological Department (IMD) currently makes annual forecasts of spatial mean rainfall over summer monsoon months of July-September (JJAS), called Indian Summer Monsoon Rainfall (ISMR). We define an analogous quantity for the entire year, All-India Mean Rainfall (AIMR), and denote by Y_t . Its anomalies are relative to its interannual mean μ and standard deviation σ .

Large anomalies in ISMR are declared by IMD as excess or deficient rainfall years. However, rainfall is highly heterogeneous spatially. Therefore in order to define anomalies in the aggregate measure of AIMR, we consider not only calculations of Y_t but also the frequencies of local anomalies in the corresponding year.

A Markov Random Field is an undirected graph, with nodes for each (s, t) pair. Corresponding to each (s, t) pair is associated a latent variable Z_{st} and an observation Y_{st} . Nodes corresponding to (s, t) and (s', t) possess an edge between them if s and s' are neighbouring grid-points (each location has 8 neighbours, except border locations), for each year t . Likewise, for each location s , nodes corresponding to neighbouring years, i.e. (s, t) and $(s, t+1)$, have an edge. Each observation node $Y(s, t)$ has a single edge, to the corresponding latent variable node $Z(s, t)$.

The graph also contains nodes corresponding to each year, associated with latent Z_t and observed Y_t , corresponding to AIMR. For any year t , Z_t is linked by edges to all nodes for that year $\{Z_{st}\}$ for every location s .

Probabilities are assigned to each configuration of Z using node *potential functions* $\psi^v(Z_{st})$ on each node, edge potentials $\psi^e(Z_{st}, Z_{s't'})$ on each edge occurring between spatiotemporal nodes and $\psi^f(Z_{st}, Z_t)$ on each edge occurring between spatiotemporal nodes and AIMR nodes. Edge potentials influence spatial and temporal coherence and node potentials influence the threshold for anomaly detection. As will become clear from the Gibbs sampling equation in Section 2.5, the node potential functions can be interpreted as describing the

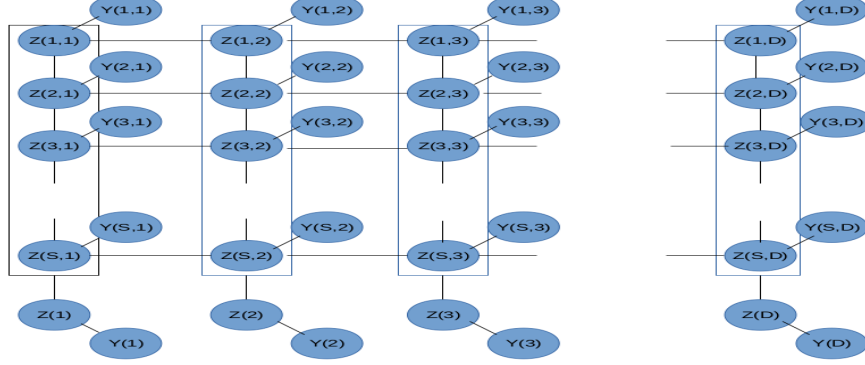


Figure 1: Proposed Markov Random Field for Anomaly Detection

prior probability distribution across different states. Edge potentials describe prior probabilities that the nodes connected by the edge are in the same state.

The precipitation amount at any location and year, given by Y_{st} , is modelled using a Gaussian distribution with parameters specific to the location s and latent state Z_{st} . One interpretation of such a conditional distribution is as edge potentials between the $Z_{st} - Y_{st}$ edges connecting the latent and observed states respectively.

The likelihood function is:

$$L(Z) \propto \prod_{s,t} \psi^v(Z_{st}) \prod_e \psi^e(Z_{st}, Z_{s't'}) \prod_f \psi^f(Z_{st}, Z_t) \prod_{s,t} \mathcal{N}(Y_{st}; \mu_{sZ_{st}}, \sigma_s) \prod_t \mathcal{N}(Y_t; \mu_{Z_t}, \sigma) \quad (1)$$

This defines the likelihood function, i.e. the probability of observing the data given the latent variables in the graph.

2.4 Spatial and Temporal Coherence through MRF

The spatiotemporal rainfall volume Y_{st} is modeled as a multi-modal Gaussian distribution, and Z_{st} specifies the mode. The parameters (μ_{sp}, σ_s) of this distribution depend on the latent state p as well as location s , and are estimated from data. Similarly for spatial mean rainfall Y_t we use a Gaussian distribution with state-specific parameters (μ_p, σ) . Initial estimates of these parameters can be made from the dataset using LWA to assign states.

We define **edge potential functions** so that if two vertices connected by an edge have same values of Z then the corresponding edge potential is larger than if the values were different. Since the likelihood function is multiplied by these edge potentials, this encourages spatial and temporal neighbours to have same state, leading to spatial and temporal coherence. For each edge between location state node Z_{st} and the corresponding AIMR state node Z_t for the same year, the edge potential influences the extent to which the local state is sought to be made coherent with the aggregate state. We define potential functions for

different edges as follows:

$$\begin{aligned}
\psi^e(Z_{st} = Z_{s't}) &= C(s, s'), \psi^e(Z_{st} \neq Z_{s't}) = D \\
\psi^e(Z_{st} = Z_{s,t+1}) &= P, \psi^e(Z_{st} \neq Z_{s,t+1}) = 1 - P \\
\psi^f(Z_{st} = Z_t) &= \exp(1/S), \psi^e(Z_{st} \neq Z_t) = 1
\end{aligned} \tag{2}$$

To emphasize spatial coherence, D is a small constant compared to $C(s, s')$. The latter describes edge potentials if spatial neighbours are in the same state. As described previously, these edge potentials can be viewed as prior probabilities on the neighbours being in the same state. Therefore $C(s, s')$ represents a prior probability that the states in locations s and s' are the same, and is estimated from data. Two neighbouring grid-locations need not be highly correlated, for e.g. on either side of a narrow mountain range (such as the Western Ghats). Therefore unlike the MRF estimated by [4], where all edges between neighbouring pairs have the same potential function, here the potentials on edges are estimated from data and are location-dependent.

The value of edge potential P , for edges connecting nodes with neighbouring years, lies between zero and one. It induces temporal coherence, and hence is called the temporal coherence parameter. Higher values induce a higher emphasis on temporal coherence.

The third set of edge potentials describes behaviour of edges between the location nodes in any given year and the spatial mean node for that year. It is defined using the exponential, so that the contribution depends on the total number of locations whose states coincide with the state assigned to the spatial mean node. S is the total number of locations. The edge potential is higher when the location nodes are in the same state as the spatial mean node.

Next, we define the **node potential functions**. These are directly proportional to the prior probabilities of the nodes being in the different states, and generally influence the threshold for anomaly detection in most real situations when data is finite and the prior is not immaterial in the MAP solution. The state that is eventually assigned in the MAP solution depends only on the relative values of these node potentials. For the default model, all node potentials are set equal to the same value, which is set to 1. But they can be varied according to the problem of interest, as described further in the Appendix.

MRF parameter settings: Only the part of the likelihood function that varies with the state Z affects the MAP solution. Therefore a node or edge potential can be made irrelevant to the particular analysis by making it constant, independent of the value of Z . In the subsequent sections, we will use this device to consider alternate settings of the MRF, including where either spatial or temporal coherence are considered in isolation.

2.5 Anomaly Detection by Markov Random Fields

Having defined the likelihood function, we carry out inference on the latent variables Z and estimate parameters $(\mu_{sp}, \sigma_s, C(s, s'))$ for locations s , corre-

sponding neighbours s' and conditioned on latent state p . Unlike the maximum likelihood estimation of [4] that is based on integer programming, here we carry out inference by Gibbs Sampling, which is computationally simpler [10].

Each latent variable Z_{st} is initialized based on location-wise analysis described earlier, and corresponding parameters are estimated. The Gibbs sampling technique entails, at each iteration, sampling each Z_{st} -variable from its updated conditional distribution by conditioning on values of other variables estimated thus far in the iteration, and then re-estimating the parameters. The procedure is repeated for several iterations, and samples are collected at regular intervals. The stationary distribution of this Markov chain Monte Carlo procedure is the posterior distribution on the latent variables. The maximum a-posteriori (MAP) estimate of Z -variables can then be made from the samples.

The Gibbs Sampling equation for any latent variable Z_{st} is given by:

$$\begin{aligned} p(Z_{st} = p | Z_{-s,-t}, Z_t, Y_{st}) &\propto p(Z_{st} = p, Z_{-s,-t}, Z_t, Y_{st}) \propto p(Z_{st} = p, Z_{s't}, Z_{s't'}, Z_t) p(Y_{st} | Z_{st}) \\ &\propto \psi^v(Z_{st} = p) \psi^f(p, Z_t) \prod_{s', t'} \psi^e(p, Z_{s't'}) \mathcal{N}(\mu_{sp}, \sigma_s) \end{aligned}$$

where s' refers to neighbours of s , t' to the previous and next years, i.e. $(t-1)$ and $(t+1)$ and the state $p \in \{1, 2, 3\}$. $Z_{-s,-t}$ means all the Z -variables except Z_{st} . While applying this equation, we do not consider variables corresponding to spatiotemporal locations that are not neighbours of Z_{st} , since the Markov property of MRF holds that each node is conditionally independent of all non-neighbouring nodes conditioned on the neighbouring nodes. After each step, edge potentials are estimated from equation (2). The Gibbs Sampling proceeds by drawing samples for each Z_{st} from the above equation, and the optimal value for each latent variable is estimated from the distribution across these samples.

After estimating the latent-variable-set Z , we identify anomalies by discovering spatially and/or temporally coherent clusters of spatiotemporal locations. Spatiotemporal anomalies are estimated as connected components of the MRF, such that each node of the connected component has the same state. These values of Z can be either 1 or 2, corresponding to high and low anomalies respectively. Connected components are computed using the OPTICS algorithm [3]. Due to coherence, the clusters thus identified can be at a single location but extending over several contiguous years, or spatially extended locations in a single year, or both.

2.6 Statistics of Anomalies

A number of statistics are defined for these anomalies. $NI1$ and $NI2$ are respectively the number of years of positive and negative anomalies at all-India level, satisfying $Z_t = 1$ or $Z_t = 2$. $N1$, $N2$ are the total number of spatiotemporal nodes assigned to states 1 and 2 respectively, i.e. $Z_{st} = 1$ or $Z_{st} = 2$.

The number of positive (NP) and negative (NN) anomalies, in each parameter setting, is estimated as the number of spatiotemporally connected components, each component consisting of spatiotemporal locations in the corre-

sponding state. While computing NP and NN we consider only those connected components with size larger than 1. For this comparison, we consider the spatiotemporal size defined below.

The *spatiotemporal size* of each anomaly is the size of this connected component, i.e. the number of nodes present in it. We measure the STS: mean spatiotemporal size of all anomalies, including all years; and similarly the STSP: mean spatiotemporal size of all positive anomalies; and STSN: mean spatiotemporal size of all negative anomalies. We define the *spatial size* of an anomaly as the number of distinct spatial locations included in the nodes covered by it. The *temporal size* of an anomaly is similarly defined as the number of distinct years included in it. We thereby estimate mean spatial size of all anomalies (SS), only positive (SSP) and only negative (SSN) anomalies. Similarly we measure (TS, TSP, TSN) for corresponding mean temporal sizes.

We have considered various settings of the MRF for rainfall anomaly detection, starting with the baseline Location-Wise Analysis (LWA) in Section 2.2. For the MRF-based methods, we consider variants with only spatial coherence (MRF-SC), only with temporal coherence with the temporal coherence parameter P set to p (MRF-TC- p), and MRF with both spatial and temporal coherence as $MRF - STC - x$ where x may refer to the temporal coherence parameter P or a choice of spatial coherence parameters C and D . If no parameter is mentioned, it refers to MRF with both spatial and temporal coherence, with $P = 0.99$ and the spatial coherence parameter $C(s, s')$ between pairs of locations (s, s') proportional to the number of years that they are in same "phase", i.e. both have a rise (or fall) in annual rainfall volume compared to previous year.

In comparing alternate methods for anomaly detection, one comparison to be made is the number of spatiotemporal locations assigned to an "anomaly" state ($Z_{st} = 1$ or $Z_{st} = 2$) that are either gained or lost in one method relative to another. These counts are denoted $NG1$ (gain in the number of spatiotemporal locations in state 1), and $NL1$ (loss in the number of spatiotemporal locations in state 1). Similarly for state 2 we have $NG2$ and $NL2$.

For readability, these notations are listed in Table 1.

3 Test of Method

The results from the MRF are compared with location-wise analysis (LWA) discussed previously in Section 2.1.

3.1 Comparison with LWA for given years

We examine results from two years: 1998 (declared excess-rainfall year by IMD) and 2002 (declared deficient-rainfall year). Maps of positive and negative anomalies in these two years are shown in Figure 2. The first panel in each pair shows results from the LWA, while the second panel shows those of the MRF. Overall the maps have many similarities, which seem to validate the MRF approach.

Abbreviation	Statistic Description
H	years when all-India rainfall is excess
L	years when all-India rainfall is deficient
NI1	#all-India positive anomalies
NI2	#all-India negative anomalies
N1	#spatiotemporal locations assigned state 1
N2	#spatiotemporal locations assigned state 2
N1	mean #spatiotemporal locations assigned state 1 per year
N2	mean #spatiotemporal locations assigned state 2 per year
N1H	mean #spatiotemporal locations assigned state 1 in years of H
N2L	mean #spatiotemporal locations assigned state 2 in years of L
NG1	#S-T locations in state 1 "gained" in one method over another
NG2	#S-T locations in state 2 "gained" in one method over another
NL1	#S-T locations in state 1 "lost" in one method over another
NL2	#S-T locations in state 2 "lost" in one method over another
NP	#positive anomalies found by running OPTICS
NN	#negative anomalies found by running OPTICS
SS	mean spatial size of anomalies
SSP	mean spatial size of positive anomalies
SSN	mean spatial size of negative anomalies
TS	mean temporal size of anomalies
TSP	mean temporal size of positive anomalies
TSN	mean temporal size of negative anomalies
STS	mean spatiotemporal size of anomalies
STSP	mean spatiotemporal size of positive anomalies
STSN	mean spatiotemporal size of negative anomalies

Table 1: Abbreviations for different statistics of spatiotemporal anomalies

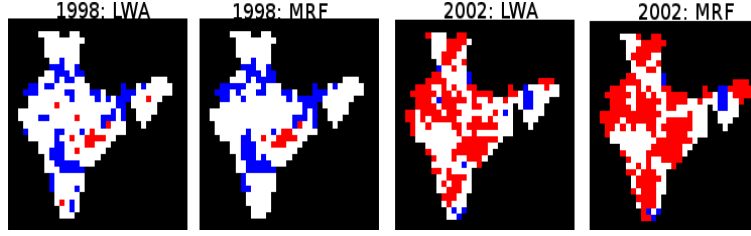


Figure 2: Comparison of results of MRF with location-wise analysis (LWA) for 1998 (excess-rainfall year) and 2002 (deficient-rainfall year). First panel in each pair shows results for LWA. Colors indicate different latent-states (blue: positive; red: negative; white: normal). In case of MRF, anomalies are more spatially contiguous.

It was noted previously that LWA may yield isolated anomalies as well, and this is seen in the figure. By contrast, the constraint of spatial coherence in the MRF yields more spatially connected and extensive anomalies, with fewer isolated anomalies. Anomalies of both kinds are more spatially contiguous with the MRF.

Furthermore, for the excess rainfall year, the MRF yields a larger number of locations with positive anomaly state ($N1$) compared to LWA (84 in MRF as compared to 75 in LWA). Likewise, in the deficit-rainfall year, the MRF yields more locations having negative anomalies ($N2$) (194 in MRF compared to 147 in LWA). This is a result of the edges connecting the location nodes to the spatial mean node in the MRF, which have higher edge potential when the corresponding nodes are in the same state, as well as the effects of spatial coherence.

3.2 Statistics of anomalies across years

The variables Z_t denote the anomaly corresponding to All-India spatial mean rainfall (AIMR). For each year we compute AIMR Y_t from local measurements $\{Y_{st}\}$, and from this time-series estimate mean μ and standard deviation σ across years. The excess rainfall years H are defined as those with $Y_t \geq \mu + \sigma$ and deficient-rainfall years L have $Y_t \leq \mu - \sigma$.

These definitions do not depend on how widespread are local anomalies but only on amount of spatial mean rainfall. We can instead define all-India anomalies so as to depend on the widespread occurrence of local anomalies. For any year t , we compute the number of local anomalies of both kinds ($N1(t)$ and $N2(t)$) as found by LWA, and corresponding means (μ_{N1} , μ_{N2}) and standard deviations (σ_{N1} , σ_{N2}) across time. Based on these, we identify those years with exceptionally large numbers of local positive anomalies (HL) and exceptionally large numbers of local negative anomalies (LL). In other words, $HL = \{t : N1(t) \geq \mu_{N1} + \sigma_{N1}\}$ and $LL = \{t : N2(t) \geq \mu_{N2} + \sigma_{N2}\}$. It turns out that H and HL are not equal, and their overlap $\frac{|\text{intersect}(H, HL)|}{|H|}$ is only 0.7. Similarly

L and LL are also not equal, and $\frac{|\text{intersect}(L,LL)|}{|L|}$ is only 0.7. This illustrates that the aggregate state Z_t when defined based on spatial mean rainfall often takes different values from when it is defined based on widespread occurrence of local anomalies.

In the MRF model, edge potentials ensure that assignment of Z_t is also influenced by values of the location-wise latent states Z_{st} , and large numbers of local anomalies of one kind increase the probability of Z_t being assigned to the same anomaly. At the same time, it also takes into account the AIMR estimate Y_t . Hence in the MRF the value of Z_t should be able to capture all kinds of all-India anomalies defined so far - H, HL, L, LL .

Let ZH and ZL be the positive and negative years identified by the MRF, i.e. $ZH = \{t : Z_t = 1\}$ and $ZL = \{t : Z_t = 2\}$. The set ZH captures very well the contents of both H and HL , with $\frac{|\text{intersect}(H,ZH)|}{|H|} = 1$ and $\frac{|\text{intersect}(HL,ZH)|}{|HL|} = 0.92$. Similarly ZL also overlaps well with L and LL , with $\frac{|\text{intersect}(L,ZL)|}{|L|} = 1$ and $\frac{|\text{intersect}(LL,ZL)|}{|LL|} = 0.84$. This shows that the MRF model helps discover both types of all-India anomalies, based on spatial-mean rainfall as well as widespread occurrence of local anomalies, simultaneously.

4 Effects of MRF Parameters on Detected Anomalies

As mentioned in the Introduction, anomaly detection is inherently subjective. The results of an MRF-based approach depend on the values of node and edge potentials. In this section, we study and evaluate effects of these parameters on the inferred latent states Z , and thereby on the properties of anomalies detected using this approach. Only that part of the likelihood function that varies with the latent state can influence the MAP solution.

4.1 Effects of Spatial Coherence

Let us isolate effects of spatial coherence, in the absence of temporal coherence. Absence of temporal coherence is implemented by using constant edge potentials for all edges across years. We also use constant node potentials for all nodes and states.

We describe anomaly statistics from the MRF for extreme years, where all-India rainfall is either excess or deficient. Generally, across approaches it can be expected that in years of H (excess rainfall) the number of locations (N1H) assigned as positive anomaly $Z_{st} = 1$ is much higher than positive anomaly locations in all other years (N1Y), while the number of locations (N2L) assigned to negative anomaly in L (deficit rainfall years) is much higher than in all other years (N2Y). These relationships are seen for the MRF with spatial coherence and LWA in Table 2.

Spatial coherence in the MRF causes the mean number of nodes with positive anomalies in years of excess rainfall to be higher than in case of LWA (Table

Method	N1Y	N2Y	N1H	N2L	D12H	D21L
LWA	54	54	107	111	101	86
MRF	62	58	132	129	118	103

Table 2: Mean number of spatial locations with positive (1) and negative (2) anomalies in all years, only excess-rain (H) years and only deficient-rain (L) years. Also, difference between number of nodes with positive and negative anomalies in H and L years. Results are shown for the MRF and location-wise analysis (LWA). Compared to LWA, spatial coherence in the MRF increases occurrence of corresponding anomaly states in excess and deficit rainfall years.

2). Similarly there are more negative anomalies in years of deficit rainfall as compared to LWA. Furthermore, the mean difference between number of nodes with positive and negative anomalies in H and L years respectively (D12H, D21L) is more pronounced with the MRF than in case of location wise analysis (Table 2). Spatial coherence favours occurrence of the corresponding anomaly states in excess or deficit rainfall years.

4.2 Effects of Temporal Coherence

Here we describe effects of temporal coherence alone. For different parameter settings, we compute the total number of nodes in the entire graph assigned states 1 and 2 ($N1$, $N2$). We also compute *confusion matrices* to describe the degree of overlap between anomaly nodes found by location-wise analysis and the MRF. $NG1$ denotes the number of positive anomaly nodes "gained" by the proposed method when compared to LWA, i.e. nodes satisfying $Z_{st} = 1, Z0_{st} \neq 1$ (Recall that state assignments by LWA are $Z0$). These are nodes not classified as positive anomalies by LWA, but that are positive anomalies in the corresponding MRF. Similarly $NL1$ is the number of positive anomaly nodes "lost" by the proposed method compared to LWA, i.e. nodes satisfying $Z0_{st} = 1, Z_{st} \neq 1$. The number of negative anomaly nodes "gained" and "lost" in this way are denoted as $NG2$ and $NL2$ respectively.

We consider effects of temporal coherence with parameter P ($MRF - TC - P$), with increasing P denoting increasing emphasis on temporal coherence. Spatial coherence is absent in this subsection. The node potential is uniform, independent of the assignment of latent variable Z . Results are shown in Table 3 and Figure 3.

In the presence of temporal coherence, the number of nodes with positive anomaly is much larger than that of locations with negative anomaly. The relative difference increases as the temporal coherence parameter increases.

As the role of temporal coherence is increased, by increasing P from 0.5 to 0.99, the number of anomalies decreases. Increasing coherence generally leads to fewer anomalies. That is why it is not possible to generalize the effect of MRF compared to LWA, without also specifying the coherence parameters.

In general the number of anomalies "lost" when switching from LWA to the

MRF is higher as either spatial or temporal coherence is introduced, and as the temporal coherence parameter is increased. This is expected, as a many anomalies found by LWA are isolated and do not reflect coherent effects on larger scales. A less expected effect of introducing coherence is that a significant number of new anomalies are "gained", i.e. identified when LWA could not extract them. Such anomalies are manifested at larger scales only.

4.3 Presence of Temporal and Spatial Coherence

Here we consider the MRF where both spatial coherence and temporal coherence, the latter having parameter P , are present ($MRF - STC - P$). Results are shown in Table 3. In the presence of spatial coherence, the effects of increasing the temporal coherence parameter P are similar to the previous discussion in the context of temporal coherence alone: higher temporal coherence parameter leads to fewer anomalies. Furthermore, the number of positive anomalies is larger than the number of negative anomalies, and the relative difference becomes larger as temporal coherence is increased.

There can be different approaches to enforcing spatial coherence, and we consider the effects in the following. We contrast five different approaches.

In the first three below, $D = 0$. That is, the edge potentials for spatial neighbours have zero weight if the latent states differ. These approaches differ in the choice of edge potentials $C(s, s')$ between spatial neighbours in case the latent states are the same: "prop", where for neighbouring pairs of locations, $C(s, s')$ is proportional to the number of years that the locations have the same "phase" i.e. sign of rainfall change; "anml" where for neighbouring pairs of locations, $C(s, s')$ is proportional to the number of years that the locations had the same anomaly as estimated by LWA; and "unif" where for neighbouring pairs of locations, $C(s, s')$ values are equal. An important result is that these approaches do not have much effect on anomaly statistics (Table 3). Therefore anomaly detection using MRFs does not depend much on details of the spatial coherence model as long as the edge potentials in the presence of spatial coherence are much higher than edge potentials when the neighbouring states differ; recall that for these three cases $D = 0$ so that ratio C/D is infinity.

In the last two approaches towards spatial coherence, we relax the constraint that $D = 0$. This is essentially a weakening of the spatial coherence requirement. Increasing the values of C and D by a constant factor affects all possible assignments of latent states equally, and cannot influence the solution. The ratio of C and D can, however, affect the relative weight given to spatial coherence, with higher ratios emphasizing spatial coherence more. We consider two settings: "mxd1" where $C = 2, D = 1$ and "mxd2" where $C = 5, D = 1$. If ratio C/D is higher, there are fewer anomaly nodes (Table 3).

Method	N1	N2	NG1	NG2	NL1	NL2
LWA	5666	5621	-	-	-	-
MRF-SC	6561	6038	481	678	1376	1095
MRF-TC-0.50	7905	7645	2248	2031	9	7
MRF-TC-0.75	6687	6379	1319	1079	298	321
MRF-TC-0.90	5482	4725	1065	725	1249	1621
MRF-TC-0.99	3555	2178	910	408	3028	3844
MRF-STC-0.50	6484	6049	1313	1090	495	663
MRF-STC-0.75	4916	4322	583	410	1333	1709
MRF-STC-0.90	3447	2282	361	185	2580	3524
MRF-STC-0.99	1828	1105	204	96	4042	4612
MRF-STC-unif	1755	1013	192	83	4103	4691
MRF-STC-prop	1828	1105	204	96	4042	4612
MRF-STC-anml	1808	1109	196	113	4054	4625
MRF-STC-mxd1	3125	1785	704	276	3252	4105
MRF-STC-mxd2	2200	1328	379	166	3934	4597

Table 3: Total number of nodes assigned to the different anomaly states of the entire graph, in different settings of MRF, and the number of anomaly nodes "gained" and "lost" compared to location-wise analysis (LWA). See Section 4.2 for definitions. Increasing the temporal coherence parameter decreases the number of anomaly nodes. Increasing the temporal coherence parameter makes positive anomaly nodes more predominant. Increasing the ratio of C and D in the spatial coherence model leads to fewer anomaly nodes.

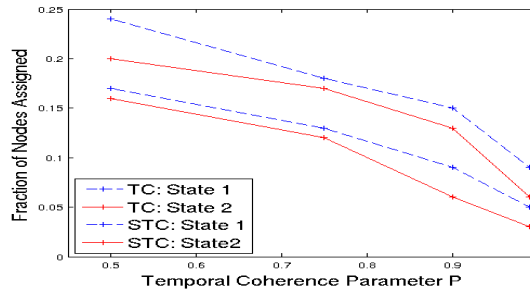


Figure 3: Fraction of spatiotemporal locations assigned to both anomaly states in two settings of MRF: using only temporal coherence and using both spatial and temporal coherence

5 Spatial and Temporal Scales of Anomalies

An important question is how widespread and persistent positive and negative rainfall anomalies are. Following the definitions in Section 2.6, we examine the influences on the spatiotemporal, spatial, and temporal sizes of positive and negative anomalies.

5.1 Effects of edge potentials

We consider location-wise analysis (LWA), and using MRFs under different settings. These settings include only spatial coherence (SC), only temporal coherence with parameter P (TC_P) and both (STC_P). Results are shown in Table 4. The different groups of columns show the number of anomalies, spatiotemporal size, spatial size, and temporal size, respectively.

The results indicate complex relationships involving spatial and temporal scales of anomalies. As expected, with LWA, the number of anomalies is much larger and their mean sizes much smaller, in comparison to versions of the MRF where various constraints of coherence are present.

In the presence of temporal coherence, as the temporal coherence parameter is increased, the spatial size of anomalies becomes smaller. Larger temporal coherence parameter selects for more long-lived anomalies and hence these tend to become smaller in spatial extent. The spatiotemporal size decreases as the temporal coherence parameter is increased.

The aforementioned effect is also present when spatial coherence is included in the MRF. The selection for longer but spatially less extended anomalies when the temporal coherence parameter is increased creates a trade-off between spatial and temporal extents. Such a trade-off is intrinsic to anomaly detection, and is made explicit in Figure 4. In Figure 4, the mean temporal and mean spatial sizes of anomalies are compared in different settings, obtained by varying parameter P that prescribes the edge potential in the presence of temporal coherence. Larger mean spatial size corresponds to shorter mean temporal size.

When spatial coherence is introduced in addition to temporal coherence, the spatial size of anomalies increases and the temporal size decreases. There is a general lesson about coherence from these results: with a larger emphasis on a certain type of coherence (spatial or temporal) the corresponding size of anomalies increases while the other size decreases.

The last set of results in Table 4 describes effects of changing the edge potentials enforcing spatial coherence. We note the previous discussion of Section 4.3, where we considered the total number of nodes assigned to different anomaly states under different settings of these parameters. The first three settings have $D = 0$, or zero edge potential between spatial neighbours if latent states differ. A result of Section 4.3 was that alternate approaches for choosing C in this case do not significantly influence the number of nodes assigned to anomalies. Table 4 shows that spatial, temporal, and spatiotemporal size are not sensitive to the strategy for choosing C , if $D = 0$.

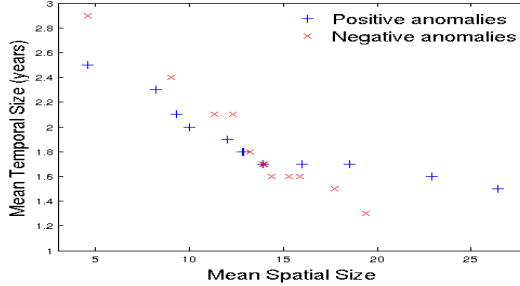


Figure 4: Trade-off between mean spatial size and mean temporal size of positive and negative anomalies, by varying temporal coherence parameter P . Spatial coherence is present in this model, and node potentials are uniform.

The last two rows of Table 4 involve relaxing the constraint that $D = 0$, so that edge potentials for spatial neighbours can be non-zero even while latent states differ. In case of lower ratio between C and D , emphasizing spatial coherence less, there are more anomalies, anomalies with smaller spatial size are detected, and correspondingly the mean temporal size is longer.

The above discussion pertained to parameter-based tradeoffs in mean spatial and temporal sizes of anomalies. However, even for fixed parameter settings of the MRF, there is substantial variation in anomaly sizes. Such variation of spatial and temporal sizes is shown in Figure 5 for one realization of the MRF. It is seen that larger anomalies tend to be shorter-lived, but there are individual exceptions. Knowing the spatial size of an anomaly can reduce uncertainty about the temporal extent for which the anomaly might persist, but such constraints are likely to be weak. There is a large range of temporal sizes for a known spatial size, for both positive and negative anomalies.

5.2 Effects of node potentials

Node potentials influence the thresholds for anomaly detection, and can be interpreted as prior probabilities of the corresponding anomaly being present before any observations are made. To examine the effects we compute the *mean number* of positive (NP) and negative (NN) anomalies with spatiotemporal size above 1, as well as their mean spatiotemporal sizes, all defined in Section 2.6. In all cases, we maintain spatial and temporal coherence through edge potentials in the MRF, with temporal coherence parameter $P = 0.99$.

In setting NP1, we consider equal weights for all 3 states at each node; NP2 favours detection of positive anomalies by setting $C_1 = 2, C_2 = 1, C_3 = 1$; NP3 favours negative anomalies by setting $C_1 = 1, C_2 = 2, C_3 = 1$ in NP3; NP4 prioritizes both anomalies over the normal state using $C_1 = 2, C_2 = 2, C_3 = 1$.

One might also set node-specific potentials depending on statistics at either the location or the year associated with the node. We define set LS of dry locations, where mean annual rainfall (μ_s) is at least one standard deviation σ

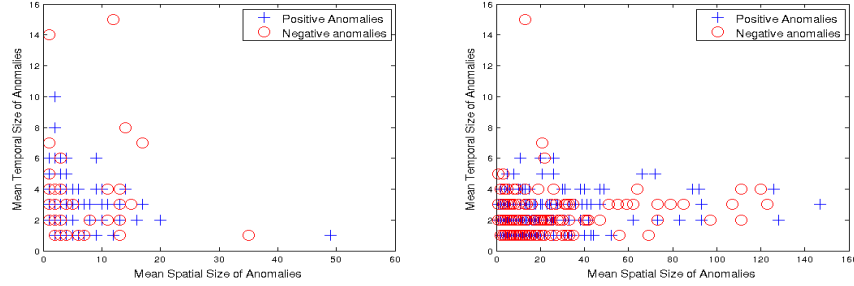


Figure 5: Temporal versus spatial sizes of individual positive and negative anomalies, in fixed parameter settings. Spatial coherence (prop) is used with two choices of the temporal coherence parameter (left: $P = 0.99$, right: $P = 0.50$) and uniform node potentials. Larger anomalies tend to be shorter-lived, but there are individual exceptions and large variability exists in the sizes of individual anomalies.

Method	#Anomalies		Spatiotemporal size		Spatial sizes		Temporal sizes	
	NP	NN	STSP	STSN	SSP	SSN	TSP	TSN
LWA	572	513	9.3	10.2	7.3	7.8	1.8	1.9
MRF-SC	392	336	16.5	17.8	12.8	13.9	1.8	1.7
MRF-TC-0.50	630	625	12.0	11.7	8.7	8.4	2.2	2.2
MRF-TC-0.75	585	588	11.0	10.4	6.8	6.6	2.6	2.4
MRF-TC-0.90	528	507	10.0	8.9	5.2	4.7	3.0	2.8
MRF-TC-0.99	371	293	9.3	7.3	3.1	2.2	4.3	3.5
MRF-STC-0.50	384	333	16.7	17.9	12.9	14.0	1.8	1.7
MRF-STC-0.75	331	227	14.4	18.2	10.0	12.3	2.0	2.1
MRF-STC-0.90	250	135	13.3	16.6	8.2	9.0	2.3	2.4
MRF-STC-0.99	185	76	9.4	13.9	4.6	4.6	2.5	2.9
MRF-STC-unif	176	73	9.7	14.2	4.7	4.7	2.6	2.9
MRF-STC-prop	185	76	9.4	13.9	4.6	4.6	2.5	2.9
MRF-STC-anml	183	77	9.7	14.2	4.6	4.6	2.5	3.0
MRF-STC-mxd1	347	223	8.3	7.7	3.2	2.5	4.2	3.6
MRF-STC-mxd2	293	158	7.5	8.6	3.0	2.7	3.5	3.5

Table 4: Mean spatial, temporal, and spatiotemporal sizes of positive and negative anomalies in different settings of edge potentials of MRF. See Section 4.3 for definitions. A trade-off between the spatial and temporal sizes of anomalies is inherent to anomaly detection; and illustrated here by varying the temporal coherence parameter. Larger spatial coherence effect in the MRF leads to larger spatial size of detected anomalies, which correspondingly have shorter mean temporal size. Larger temporal coherence parameter leads to longer mean temporal size and correspondingly smaller mean spatial size.

below the mean of this quantity across locations (μ), i.e. $LS = \{s : \mu_s \leq \mu - \sigma\}$. We also define set HS of wet locations, where $HS = \{s : \mu_s \geq \mu + \sigma\}$.

In NP5 we set node potentials $C_1 = 2, C_2 = 1$ in nodes of HS, and $C_1 = 1, C_2 = 2$ in nodes of LS. This favours positive anomalies in wet locations, and negative anomalies in dry locations. In contrast, the values are reversed in NP6, favouring positive anomalies in dry locations and negative anomalies in wet locations.

For introducing year-specific node potentials, we consider deficient-rain years L and excess-rain years H once again. In NP7 we set $C_1 = 2, C_2 = 1$ in nodes of H, and $C_1 = 1, C_2 = 2$ in nodes of L. This favours positive anomalies in excess-rain years and negative anomalies in deficit-rain years. These settings are reversed in NP8, favouring positive anomalies in deficit-rain years and negative anomalies in excess-rain years.

Table 5 shows anomaly statistics for the various settings of node potentials examined here. When giving additional weight to positive anomalies (as in cases NP2, NP4) the number of positive anomalies increases as would be expected. Similarly when negative anomalies are given higher weight (as in cases NP2, NP4) the number of negative anomalies increases. A common tendency across these settings is that the number of distinct positive anomalies is much larger than that of negative anomalies, but negative anomalies have larger mean spatiotemporal size.

Emphasizing node-specific potentials that depend on features of either the location or the year associated with the node, in NP5-NP8, does not substantially change the overall statistics, but affects the particular anomalies detected (which are not shown). In NP7, where in AIMR anomaly years the local anomalies of the same type are favoured, the difference between mean sizes of negative and positive anomalies decreases. This is mainly because positive anomalies have higher spatial size than negative anomalies in this condition. The aforementioned situation is reversed in NP8, when in the anomaly years local anomalies of the reverse type are favoured.

6 Discovery of Nonstationarities in Anomalies

6.1 Locations of Change

Previous studies have shown nonstationarity in rainfall on daily timescales during 1951-2000, with increasing frequency of heavy rainfall events combined with decrease in moderate rainfall events, to yield a nearly stationary time-series of rainfall during June-September (Goswami et al. [19]). The anomalies considered in the present work are with respect to annual means of rainfall, and not directly relevant to the consideration of individual extreme events. The spatially and temporally extended anomalies considered here, especially extended droughts, can have substantial effects especially if occurring in areas vulnerable to rainfall variability.

This section considers changes in the frequencies of anomalies over time,

Method	NP	NN	STSP	STSN	SSP	SSN	TSP	TSN
NP1	185	76	9.4	13.9	4.6	4.6	2.5	2.9
NP2	211	83	11.5	12.8	4.9	4.5	2.8	2.7
NP3	185	111	9.9	12.9	4.6	4.4	2.6	3.4
NP4	206	116	11.7	13.8	5.1	4.4	2.8	3.0
NP5	186	75	10.1	14.1	4.6	4.7	2.7	2.9
NP6	188	88	9.8	13.3	4.5	4.4	2.6	2.9
NP7	189	80	11.1	12.8	5.1	4.6	2.7	2.8
NP8	185	75	9.7	14.5	4.5	4.7	2.6	2.9

Table 5: Mean spatial, temporal, and spatiotemporal sizes of positive and negative anomalies in different settings of node potentials of the MRF. The number of distinct positive anomalies is much larger than that of negative anomalies, while negative anomalies have larger mean spatiotemporal size, across these settings. Emphasizing node-specific potentials that depend on statistics at either the location or the year associated with the node (NP5-NP8) does not substantially alter these overall statistics.

identifying significant nonstationarities present. The MRF used has uniform node potentials, with both spatial and temporal coherence and temporal coherence parameter $P = 0.99$.

Figure 6 compares the period 1901-1950 with 1951-2000, identifying locations where numbers of anomalies of each kind differ markedly (by ± 3) between the two. Upper panels show results from the MRF. Left panels show results for positive anomalies and right panels for negative anomalies. Figure 6A shows significant changes in occurrence of positive anomalies, with increases as well as decreases occurring throughout the country. Similarly, Figure 6B shows that there have been significant changes in the frequency of negative anomalies in some locations. Bottom panels indicate corresponding results for LWA.

For the MRF (Figure 6A and 6B), the overall incidence of locations having significant changes in the frequency of anomalies is not large, but it must be recalled that anomalies are themselves rare occurrences and the baseline numbers are themselves small, especially with constraints of temporal and spatial coherence present. With LWA (Figure 6C and 6D), coherence constraints are absent, leading to a higher baseline frequency of anomalies. Hence many more locations appear as experiencing the same change in number of anomalies between the two periods compared.

The figure also shows that, for both types of anomalies, the number of locations experiencing increase and decrease are roughly similar. This suggests that there has been no large change in the frequencies of either kind of anomalies with respect to annual rainfall.

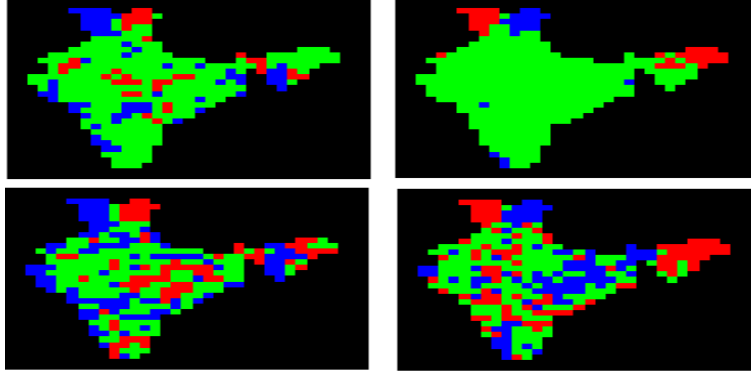


Figure 6: Change in number of anomalies of annual rainfall from 1901-1950 to 1951-2000, from both the MRF and location wise analysis. Left: Positive anomalies, Right: Negative Anomalies. Above: MRF with default parameters, Below: Location-wise Analysis. Blue: locations where number of anomalies have increased by at least 3 in second half of 20th century, Red: locations where they have decreased by at least 3. There are locations experiencing both increase and decrease in the frequency of anomalies, with neither increase nor increase dominating.

6.2 Season-wise Analysis

The above picture changes if considering rainfall received according to season, where there are distinct season-wise shifts in the frequency of anomalies. Table 6 summarizes aggregate statistics of anomalies using the MRF, in the two 50-year periods considered above.

The table reports $NI1$ and $NI2$, the number of positive and negative anomalies respectively of all-India Spatial Mean rainfall; the total number of spatiotemporal locations assigned to states 1 and 2 ($N1$ and $N2$); and the number of anomalies (NP and NN). These results are reported for three periods, delineated in the different pairs of rows: (1) the four monsoon months (JJAS) in each year; (2) pre-monsoon (April-May) and post-monsoon (October-November) months, abbreviated as AM+ON; and (3) the entire year. Each pair of rows shows statistics for the first half and second half of the 20th century.

For each statistic, the mean and standard deviations across several runs of the MRF were computed and corresponding confidence intervals are reported in the table. If these intervals for the two periods do not overlap, a significant change is noted to have occurred. For a given statistic, let (μ_1, σ_1) and (μ_2, σ_2) denote mean and standard deviations in the two periods respectively. If, for example, $\mu_1 - \sigma_1 > \mu_2 + \sigma_2$, then an increase has definitely occurred. Correspondingly, numbers listed in bold indicate those pairs showing definite changes from the first period to the second.

These results show many changes occurring when season-wise anomalies are

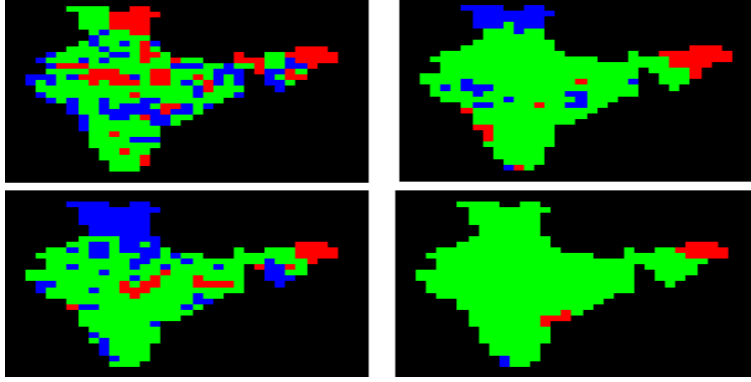


Figure 7: Change in number of anomalies with season-wise analysis from 1901-1950 to 1951-2000, from both the MRF and location wise analysis. Left: Positive anomalies, Right: Negative Anomalies. Above: Monsoon rainfall (June-September) for each year, Below: Pre-monsoon (April-May) and post-monsoon rainfall (October-November) for each year. Blue: locations where number of anomalies have increased by at least 3 in second half of 20th century, Red: locations where they have decreased by at least 3. During pre and post monsoon, there are several locations where positive anomalies have become more frequent.

considered. For the monsoon period (JJAS), there is a substantial decrease in positive anomalies of all-India rainfall, as $NI1$ is smaller in the second half. The number of distinct positive anomalies, NP also decreases, while significantly many more locations $N2$ are assigned to negative anomaly state 2. These indicate a drying trend in the monsoon months.

The tendency is the opposite when considering pre and post monsoon (AM+ON), where a general wetting trend is visible. Positive anomalies of all-Indian rainfall increase in frequency ($NI1$) in the second half of the 20th century, while negative anomalies decrease ($NI2$). More locations are assigned to positive anomaly states in the second half of the century ($N1$), while fewer are assigned to negative states ($N2$). The number of positive anomalies (NP) also increases.

The third set of results for anomalies in rainfall received in the entire year reflects a balance between these contrasting effects on JJAS and AM+ON.

Figure 7 identifies locations where there are large changes in the frequency of anomalies in the two periods. Only the MRF is used in this case. Upper panels indicate results for JJAS, while lower panels show AM+ON. During AM+ON, there are several locations where positive anomalies have become more frequent.

Table 7 lists corresponding statistics of the anomalies' sizes. Shown are mean and standard deviation of spatiotemporal size ($STSP$ and $STSN$), spatial size (SSP and SSN), and temporal size (TSP and TSN). For monsoon months, there is decrease in spatiotemporal size of positive anomalies and increase for negative anomalies. Positive anomalies have become smaller, i.e. their spatial size has decreased. Negative anomalies have become longer, i.e. their tempo-

	#All-India Anomalies		#S-T Anomaly Locations		#S-T Anomalies	
Period	NI1	NI2	N1	N2	NP	NN
1901-1950-1	11.8±0.7	10.2±1.2	783±40.1	370±8.5	62.2±2.5	19.8±4.7
1951-2000-1	7.6±0.5	10.6±0.8	763.6±16.3	637.2±25.6	92.6±2.9	27.6±2.9
1901-1950-2	5.6±0.8	14.2±1.3	388±52.7	203.8±16.9	51.2±6.9	15.2±2.3
1951-2000-2	10.8±0.4	5.8±1.2	1149.8±39.5	107.4±12.2	70.2±2.9	13±3.3
1901-1950-3	7.8±0.4	11.4±1.0	519.4±19.7	427.6±16.0	60±5.1	27.2±1.3
1951-2000-3	11.4±0.5	6.2±0.4	685.6±21.2	268.6±22.1	89.4±3.4	21.6±2.9

Table 6: Statistics of anomalies between 1901-1950 and 1951-2000 found using the MRF. Numbers of positive and negative anomalies of all-India spatial mean, total number of nodes assigned to the two anomaly states, and total number of anomalies are shown. For both periods, the comparison is done for monsoon (1), non-monsoon (2) and full years (3). The values reported are the mean over several runs of the algorithm, followed by corresponding standard deviations.

	#SpatioTemporal Size		#Spatial Size		#Temporal Size	
Period	STSP	STSN	SSP	SSN	TSP	TSN
1901-1950-1	12.38±1.0	19.96±3.9	5.38±0.2	6.88±1.2	2.54±0.2	3.76±0.6
1951-2000-1	8.2±0.3	26.74±2.4	3.84±0.2	2.64±0.0	4.92±1.8	4.98±0.5
1901-1950-2	7.06±0.7	13.14±1.5	4.24±0.3	5.36±0.7	1.84±0.2	2.24±0.3
1951-2000-2	16.2±0.9	8.64±1.7	5.22±0.1	3.1±0.5	2.78±0.2	2.72±0.2
1901-1950-3	8.36±0.4	15.92±0.7	4.12±0.2	4.94±0.4	2.3±0.1	3.44±0.2
1951-2000-3	7.6±0.3	16.58±2.2	3.14±0.1	4.16±0.4	2.92±0.0	3.6±0.2

Table 7: Statistics of anomalies between 1901-1950 and 1951-2000 found using the MRF. Mean spatiotemporal, spatial and temporal sizes of positive and negative anomalies are listed. For both periods, the comparison is done for monsoon (1), non-monsoon (2) and full years (3). The values reported are the mean over several runs of the algorithm, followed by corresponding standard deviations.

ral size has increased. All these statistics validate the broad observation that anomalies in monsoon rainfall exhibits a drying trend in the second half of 20th century, with compensating increase in rainfall during other months. This is described further in the Appendix, where Figure 8 shows the distribution of rainfall across months for the two periods.

As for pre and post monsoon (AM+ON), the reverse occurs. The spatiotemporal size of positive anomalies is larger in the second half of the 20th century, while that of negative anomalies is smaller. Positive anomalies have become larger in spatial extent while negative anomalies have become smaller. Positive anomalies have also become longer in time. This is consistent with the general wetting trend present in these months discussed with respect to Table 6.

Considering statistics based on rainfall cumulated over entire years, positive anomalies have become smaller in spatial extent but longer on average. The differences are much smaller than with the aforementioned seasons considered

separately, indicating some compensation in the effects between seasons.

This analysis brings out the following finding: monsoon rainfall in India has decreased in the second half of the 20th century, but this has been compensated by a rise in "non-seasonal" rainfall, both pre-monsoon and post-monsoon. The results are much more evident when considering the frequencies and sizes of anomalies, but a weak signal is also present in the graphs of monthly mean rainfall shown in Figure 8. As Figure 8 shows, there is a small decrease in monthly mean rainfall during JJAS that is partly offset by increase during AM+ON. As the present analysis of this section shows, such shifts in rainfall pattern have strong effects on the statistics of anomalies. Specifically, the present analysis brings out significant changes in the spatial and temporal scales of anomalies.

7 Conclusions

This paper describes a method for coherent anomaly detection using Markov Random Fields (MRFs), where each node is associated with a location and year. Coherence is emphasized because it is an inherent property of rainfall, and also because anomalies are consequential especially when extended spatially or temporally. The anomaly states are represented as latent random variables, so probabilistic methods are required for their estimation. For this purpose we use Gibbs sampling, a type of Markov chain Monte Carlo method. We also consider sensitivities of the results to parameters of the MRF.

The MRF is able to identify more coherent anomalies compared to traditional analysis using location-specific thresholds. Furthermore the method can be used to characterize both the occurrence of anomalies at large spatial scale by assigning a state variable for All-India spatial mean, as well as the widespread occurrence of grid-scale anomalies through effects of edge potentials and spatial coherence in the MRF.

The effects of edge potentials enforcing coherence as well as node potentials influencing the threshold for anomaly detection within the MRF are described. We show that adjusting the parameters has effects that are consistent with intuition. However the results are not overly sensitive to the parameters. MRFs offer one principled approach to analyzing these effects in the presence of heterogeneity and anisotropy in the occurrence of anomalies, where more traditional methods such as wavelets may not be appropriate.

One effect of coherence is to reveal anomaly states that are classified as normal in location-wise threshold-based analysis, because of the influence of neighbouring locations being assigned to anomaly states. Increasing spatial coherence through edge potentials leads to fewer but larger anomalies. Enforcing any one type of coherence more strongly, selects for either longer-lived or spatially more extended anomalies, though fewer in number.

Different settings for the edge potentials enforcing spatial coherence were considered. The results suggest that alternate strategies for choosing the edge potentials when adjacent nodes are in the same state have limited effect on the results.

Emphasizing one form of coherence over another is one way to examine the inherent tradeoff that exists between temporal and spatial scales in anomaly detection. If one scale, either temporal or spatial, is emphasized more strongly, the other scale becomes smaller.

There is also variability in the spatial and temporal sizes of anomalies. Anomalies longer in one dimension (spatial/temporal) tend to be shorter in the other. Furthermore positive anomalies are not necessarily larger or smaller than negative anomalies, as the results vary with choice of parameters.

We studied nonstationarity in anomaly statistics by comparing the first and second halves of the 20th century. Locations where numbers of positive or negative anomalies in annual rainfall changed significantly between these periods were identified. When considering annual-scale rainfall, there is no strong drying or wetting trend when comparing differences between positive and negative anomalies in the two periods.

More significant is a shift in the seasonal distribution of rainfall away from monsoon months (June-September) to other months. We consider statistics of anomalies in pre and post monsoon months (April-May and October-November), discovering a very significant rise in number as well as sizes of positive anomalies. This is compensated by decrease in number and size of positive anomalies during June-September. Similarly for negative anomalies, the frequency and scales have increased during monsoon months and decreased in other months. These effects are visible only weakly from rainfall volume data but are magnified by the lens of anomaly detection.

Overall, this study provides some understanding of heterogeneities in rainfall over Indian region. The results also raise the question of whether the anomalies discovered by this method are relevant for understanding hydrological floods and droughts, which are based on considering multiple variables, including soil moisture. A natural extension of this work would be to infer anomaly states based on the inclusion of additional climatic and hydrological variables.

Clearly, anomalies are a very significant feature of rainfall in general and Indian rainfall in particular, and any realistic simulation of regional rainfall should be able to capture their salient properties. Statistics of coherent anomalies learnt from MRF-based approaches could present further tests and benchmarks of regional-scale rainfall simulations made from climate models and statistical simulators.

8 Acknowledgments

This research was supported by Divecha Centre for Climate Change, Indian Institute of Science. We are thankful to Dr. J. Srinivasan and Dr. V.Venugopal for valuable inputs.

References

- [1] Kisilevich, Slava and Mansmann, Florian and Nanni, Mirco and Rinzivillo, Salvatore, (2009); *Spatio-temporal Clustering*
- [2] Gadgil, Sulochana and Gadgil, Siddhartha, (2006); *The Indian Monsoon, GDP and Agriculture*, Economic and Political Weekly, 4887-4895
- [3] Ankerst, Mihael and Breunig, Markus M and Kriegel, Hans-Peter and Sander, Jörg, (1999); *OPTICS: Ordering Points to Identify the Clustering Structure*; ACM Sigmod Record; Vol 28 (2); pp. 49–60
- [4] Fu,Qiang and Banerjee,Arindam and Liess,Stefan and Snyder, Peter K. (2012); *Drought detection of the last century: An MRF-based approach*; SIAM International Conference on Data Mining (SDM)
- [5] Ghosh, Subimal and Das, Debasish and Kao, Shih-Chieh and Ganguly, Au-roop R. (2012); *Lack of uniform trends but increasing spatial variability in observed Indian rainfall extremes*; Nature Climate Change, Vol. 2(2); pp 86–91
- [6] Ideiã, Sandra Maria Araújo and Santos, Celso Augusto Guimarães. (2009); *Analysis of Precipitation Time Series using the Wavelet Transform*; Revista Sociedade & Natureza; Vol 1(1)
- [7] Sharma, Aditi (2006); *Spatial data mining for drought monitoring: An approach using temporal NDVI and rainfall relationship*; International Institute for Geoinformation Science and Earth Observation, Master thesis
- [8] Narisma, Gemma T and Foley, Jonathan A and Licker, Rachel and Ramanakutty, Navin (2007); *Abrupt changes in rainfall during the twentieth century*; Geophysical Research Letters; Vol. 34(6)
- [9] Chandola, Varun and Banerjee, Arindam and Kumar, Vipin (2009); *Anomaly Detection: A Survey*; ACM computing surveys (CSUR); Vol. 41(3)
- [10] Rue, Håvard (2001); *Fast Sampling of Gaussian Markov Random Fields*; Journal of the Royal Statistical Society: Series B (Statistical Methodology); Vol. 63(2), pp 325–338
- [11] Rouault, Mathieu and Richard, Yves (2005); *Intensity and Spatial extent of Droughts in Southern Africa*; Geophysical Research Letters; Vol. 32(15)
- [12] Neal, Radford M (1993); *Probabilistic inference using Markov chain Monte Carlo methods*; Department of Computer Science, University of Toronto Toronto, Ontario, Canada
- [13] Bishop, Christopher M. (2006); *Pattern Recognition*; Machine Learning, Vol. 128

- [14] Gelfand, Alan E and Diggle, Peter and Guttorp, Peter and Fuentes, Montserrat (2010); *Handbook of spatial statistics*
- [15] Shekhar, Shashi and Jiang, Zhe and Ali, Reem Y and Eftelioglu, Emre and Tang, Xun and Gunturi, Venkata and Zhou, Xun (2015); *Spatiotemporal Data Mining: A Computational Perspective*; ISPRS International Journal of Geo-Information; Vol 4(4); pp 2306–2338
- [16] Kindermann, Ross and Snell, Laurie (1980); *Markov Random Fields and their applications*
- [17] Robert, Christian and Casella, George (2013); *Monte Carlo statistical methods*
- [18] Diaconis, Persi (2009); *The Markov Chain Monte Carlo revolution*; Vol. 46(2); pp 179–205
- [19] Goswami, Bhupendra Nath and Venugopal, V and Sengupta, D and Madhusoodanan, MS and Xavier, Prince K (2006); *Increasing trend of Extreme Rain Events over India in a warming Environment*; Science; Vol 314(5804); pp 1442–1445

9 Appendix

9.1 Choice of Node and Edge Potentials

As described in Section 2.4, node potentials can be varied depending on the problem being considered. These potentials can be viewed as prior probabilities on the occurrence of different states.

For example, a lower threshold on anomaly detection is achieved by specifying $\psi^v(Z_{st} = 1) = C_1$, $\psi^v(Z_{st} = 2) = C_2$ and $\psi^v(Z_{st} = 3) = C_3$, where C_1 and C_2 are high while C_3 is low. Relative frequencies of positive and negative anomalies can be adjusted by changing C_1 and C_2 accordingly. Another application might be to vary node potentials by location. In locations receiving low average rainfall (μ_s is small), negative anomalies may be more consequential and hence important to detect. Likewise, locations receiving higher average rainfall (μ_s is high) might be more sensitive to flooding events. We define the set of locations receiving low average rainfall as L and those receiving high average rainfall as H . Then

$$\begin{aligned}
 &\psi^v(Z_{st} = 1) = C_1, \psi^v(Z_{st} = 2) = C_2 \text{ and } \psi^v(Z_{st} = 3) = C_2 && \text{when } s \in L \\
 &\psi^v(Z_{st} = 1) = C_2, \psi^v(Z_{st} = 2) = C_1 \text{ and } \psi^v(Z_{st} = 3) = C_2 && \text{when } s \in H \\
 &\psi^v(Z_{st} = 1) = C_3, \psi^v(Z_{st} = 2) = C_3 \text{ and } \psi^v(Z_{st} = 3) = C_3 && \text{in other locations}
 \end{aligned} \tag{3}$$

To achieve the above, we specify $C_1 \leq C_2$. On the contrary, the goal may be to identify positive anomalies in dry locations, or negative anomalies in wet locations, by specifying $C_2 \leq C_1$.

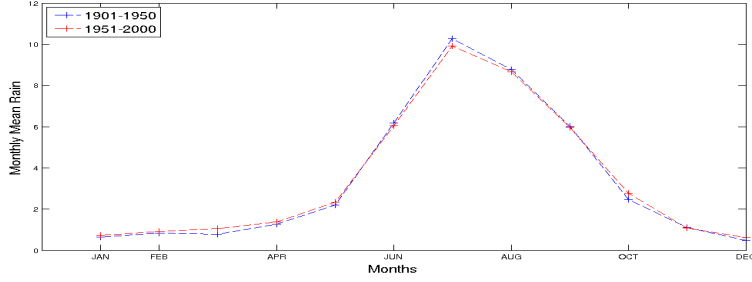


Figure 8: Seasonal distribution of rainfall averaged in the periods 1951-1950 and 1951-2000. Monsoon rainfall (June-September) decreased in the second half of the century, but increased in the other months. Pre and post monsoon rainfall (April-May and October-November) has also increased.

Yet another application may involve inducing homogeneity of heterogeneity in anomaly detection, by identifying positive anomalies especially during years of strong mean rainfall or negative anomalies in the reverse situation respectively. Alternatively, the objective may be to identify negative anomaly states during dry years or vice versa. For this type of problem, we denote sets of years with excess and deficient spatial mean rainfall as H and L . Once again defining node potentials as

$$\begin{aligned}
\psi^v(Z_{st} = 1) &= C_1, \psi^v(Z_{st} = 2) = C_2 \text{ and } \psi^v(Z_{st} = 3) = C_2 & \text{when } t \in L \\
\psi^v(Z_{st} = 1) &= C_2, \psi^v(Z_{st} = 2) = C_1 \text{ and } \psi^v(Z_{st} = 3) = C_2 & \text{when } t \in H \\
\psi^v(Z_{st} = 1) &= C_3, \psi^v(Z_{st} = 2) = C_3 \text{ and } \psi^v(Z_{st} = 3) = C_3 & \text{in other years}
\end{aligned} \tag{4}$$

Homogeneity can be achieved by specifying C_1 to be low and C_2 high, and heterogeneity with the reverse specifications. There is clear analogy between the two sets of problems, one in which node potentials are adjusted by location and the second where the type of year is the primary factor.

9.2 Seasonal Shift in Rainfall

There has been a small but noticeable shift in the seasonal distribution of spatial mean rainfall over India between the first and second halves of the 20th century. Monsoon rainfall, averaged between months of June-September, is lower during the second period. By contrast, rainfall in other months is higher for the second period. Figure 8 shows the seasonal distribution of rainfall in these two periods, where these changes are clearly visible. As shown in the main text, these changes in mean rainfall are amplified when considering the statistics of anomalies during monsoon and pre/post-monsoon months separately.